DOCUMENT RESUME

ED 182 829 EA 012 437

AUTHOR D'Elia, Michael T.
TITLE Field Test of Effective Teaching Performance Standards and Rating Scale in the Andover (Massachusetts) Public Schools. Final Report.
INSTITUTION Andover Public Schools, Mass.
PUB DATE 22 Jun 79
NOTE 25p.; For a related document, see EA 012 436

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education: *Measurement Instruments; *Performance Criteria; Records (Forms); School Surveys; *Standards; *Statistical Analysis; *Teacher Evaluation; Teachers
IDENTIFIERS *Andover Public Schools MA

ABSTRACT

This report consists of a 1979 field test of a set of ten effective teaching performance standards and a rating scale designed to evaluate teaching in the Andover Public Schools, Massachusetts. Appraisals were conducted by teams of three or four administrative evaluators, following a series of formal and informal observations. Each evaluator completed an individual report on each of 25 teachers and then participated in a composite appraisal with members of his or her team. Results showed the ten standards proved highly reliable and valid in differentiating among teachers of various abilities. Interrater reliability was also very high when measured as the degree of agreement among raters on total scores assigned teachers. The paper concludes that when the instrument is used with teachers in a regular school setting, the rating scale composed of the total ten standards will do a consistent job of differentiating teachers of various competencies. Included are copies of the performance standards and the rating scale. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED182829

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# Field Test of Effective Teaching Performance Standards and Rating Scale in the Andover (Massachusetts) Public Schools

## FINAL REPORT

Michael T. D'Elia, Ph.D.
Statistical Consultant
93 Chelmsford Street
Chelmsford, Massachusetts 01824

June 27, 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles Mitsakos

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA 012 437

TABLE OF CONTENTS


| | | Page |
|---|---|---|
| Overview | | 1 |
| I | Reliability of Instrument | 2 |
| II | Interrelater Reliability | 3 |
| III | Validities of the Standards | 4 |
| IV | Other Statistical Analysis | 4 |
| V | Conclusions | 5 |
| Appendix A | Frequency Distribution | 6 |
| Appendix B | Histogram of Total Scores Assigned | 7 |
| Appendix C | Reliability Data | 8 |
| Appendix D | Validities of the Standards | 9 |
| Appendix E | Team by Team Comparisons | 10 |
| Appendix F | Normative Data on Total Scores Assigned | 11 |
| Appendix G | Effective Teaching Performance Standards | 12 |
| Appendix H | Rating Scale | 18 |

## Overview

This is a report on the field test of a set of ten effective teaching performance standards and a rating scale designed to evaluate teaching in the Andover Public Schools. Appraisals were conducted by teams consisting of three or four evaluators (principals, program advisors, assistant principals, and/or department heads) following a series of formal and informal observations. Each evaluator completed an individual report on each teacher and then participated in a composite appraisal with members of his/her team. The assessment was conducted between February 8 and June 1, 1979. An interim report was written and presented to evaluators during April. A total of 25 teachers/specialists were evaluated.

The ten standards proved highly reliable and valid in differentiating among teachers of various abilities. Interrater reliability was also very high when measured as the degree of agreement among raters on total scores assigned teachers/specialists.

The reliability of the instrument is somewhere between .94 and .98 with a standard error of measurement of about 4 points. The validities of the standards range from a low of .61 to a high of .93, all correlating substantially with the total ratings for teachers and specialists. Nevertheless, there is a definite pattern with respect to which standards correlate highest and lowest with total scores. Things such as cognitive clarity, enthusiasm, ability to instruct diverse types of learners and providing for student interaction all correlated in the .90's while such standards as professional attitude, making use of variety, and especially standard 10, professional growth, were apparently not as important in assessing teaching ability.

Teachers were generally rated at the high end of the scale with average ratings per standard exceeding 4 on a 5 point scale. This led directly to low score variability which in turn makes

it difficult to achieve high interrater reliability when standards are considered individually. The median interrater reliability using this method was .68. However, when total scores alone are analyzed, median interrater reliability is .96, much more in line with the other reliabilities in this report. One can also hypothesize with confidence that the .68 figure would climb significantly if more poor teachers were evaluated.

The mean ratings by individual teams all reflected the "error of leniency" to some extent but the two teams who rated teachers/specialists hardest had the highest reliabilities.

## I. Reliability of the Instrument

Four reliability coefficients were calculated. The "split-half" (also known as "odd-even") reliability measures the degree to which the 5 odd-numbered standards are measuring the same thing as the 5 even-numbered standards. This was .97. Since the split-half method involves the questionable assumption that the variances of the two halves of the instrument are equal, a Guttman reliability was also calculated since it does not involve this assumption and is considered a conservative estimate. The Guttman reliability was also .97.

Kuder-Richardson reliability measures the degree to which the ten standards are measuring the same thing; hence is alternately referred to as internal consistency reliability. Formula 21, which was used, is a conservative estimate of the true internal consistency. Nevertheless, the reliability coefficient was .94.

Test-retest, or "repeat" reliability, measures the stability over time of the same raters rating the same objects. Since the elementary teams only rated the same teachers once, this reliability could only be calculated for the nine secondary teachers evaluated. It was .87 but probably would have been higher if all teachers had been included.

These coefficients compare very favorably with even the best

standardized cognitive tests, and are markedly greater than many widely used affective measures, e.g. Alport-Vernon-Lindzey "Study of Values" or "Minnesota Teacher Attitude Inventory."

Thus, the instrument itself possesses very high reliability, whatever the method used to assess it.

## II. Interrater Reliability

The results on interrater reliability depend upon one's point of view. If one considers only total scores for the ten standards assigned to teachers by different observers, the median correlation is .96, comparable to the reliabilities reported in Section I. In other words, two observers rating the same teacher/specialist would assign approximately the same total score, but one rater might be higher on some standards and lower on others than the second; these scores would tend to cancel each other out when only the total score is considered.

If, however, one wants to know how much agreement there is among raters on a standard-by-standard basis, the median reliability is .68.

The median reliabilities of the different teams varied (see appendix E). In terms of total scores, the range was from .66 to .97 with most of the reliabilities at the high end of the scale. On a standard-by-standard basis the range was from .43 to .82. The range among teams of average correlations between each pair of observers was from .29 to .68. However, if one is interested only in the total score assigned each teacher, the reliability coefficients are comparable to the instrument reliabilities reported in Section II which are very high.

The four specialists evaluated yielded interrater (total), interrater (standard-by-standard), and Pearson r reliabilities of .70, .28, and .04, respectively. Excluding them from the analysis increases all measures of interrater reliability by about .04.

Interrater Reliability (continued)

The "error of leniency" was very prevalent in this field test. Consider the fact that the median rating given every teacher on every standard was 4.13 on a 5 point scale. Evaluators willing to give both high and low ratings to teachers/specialists will obtain better reliability than evaluators who grade most teachers low or high.

## III. Validities of the Standards

The hierarchy described in the rationale and overview of "Effective Teaching Performance Standards and Rating Scale" is substantially corroborated by the results of this field test. Standards 9 and 10, at the lowest level of the hierarchy, are eighth and tenth in terms of item validities, and standard 9 only missed being ninth by .001. Moreover, standard 10 is much lower than any other standard (see Appendix D).

Perhaps the only surprises are standards 7 and 8 which ranked fifth and third, but these can only be regarded as mild surprises, since both standards have direct bearings on classroom activity, much more so than 9 and 10.

Professional growth is not a strong predictor of total scores assigned in this field test. It is the single most noteworthy finding in the validities section. At the other extreme, enthusiasm and cognitive clarity, two seemingly very unrelated traits, almost finished in a "dead heat" for first place.

## IV. Other Statistical Analysis

The mean, median, and mode for the scores of the 25 teachers/specialists observed were 120.43, 124, and 125, respectively, indicating a negatively skewed distribution wherein several low scores "pulled down" the mean relative to the median in an otherwise high-scoring distribution.

The standard deviation was 19.59, indicative of the relative lack of variability in scores, a situation which made it very

difficult to get higher standard-by-standard interrater reliability. Lower scores were used only sparingly, with almost no 1's, 1.5's, or 2's, while the 5 rating was the single most used score.

The standard error of measurement, about 4 points when teams of three evaluators are involved, is about 1.5 points when a single observer is used. One can be about 95 per cent certain in the case of a three point difference that two teachers rated by the same evaluator are truly perceived as different by the evaluator. Where an observer scores one teacher 5 points higher than another, one can be 99 per cent certain that this is not a chance occurrence.

## V. Conclusions

In reviewing the above statistical analysis, the following conclusions appear warranted:

1. In terms of the reliability of the instrument, it is clear that the ten standards yield remarkably consistent results.
2. The interrater reliability is excellent when viewed from the perspective of total scores assigned. The standard-by-standard reliability will increase to the extent that teachers of all ability levels are evaluated.
3. The validity of the individual standards when correlated with the total score for each teacher evaluated is very high with the exception of standard 10..
4. When the instrument is used with teachers in a regular school setting and thus more representative than was possible in a field test situation, the rating scale composed of the total ten standards will do a consistent and excellent job in differentiating teachers of various competencies.

## Appendix A

### Frequency Distribution

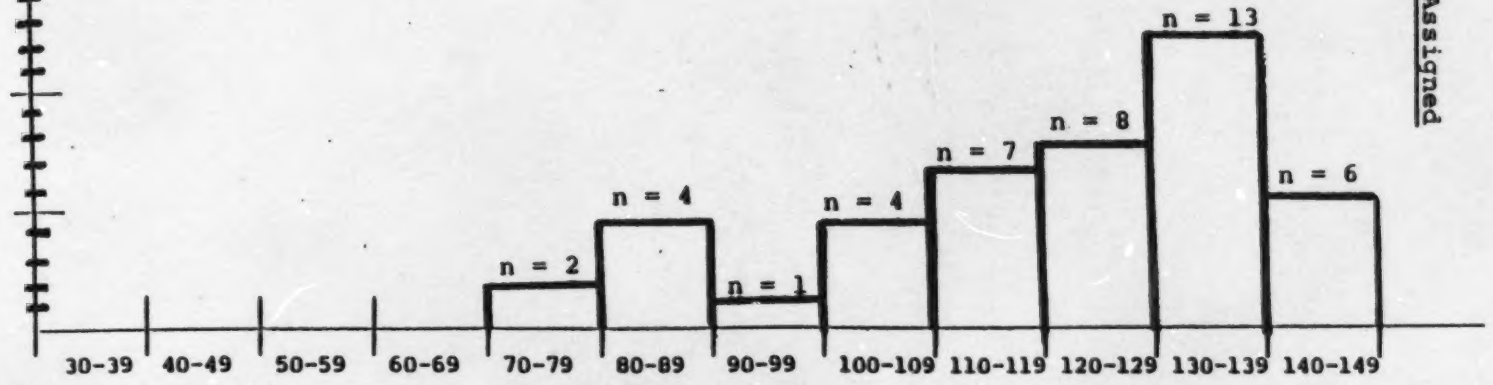
| Interval | f |
|---|---|
| 140-149 | 6 |
| 130-139 | 13 |
| 120-129 | 8 |
| 110-119 | 7 |
| 100-109 | 4 |
| 90-99 | 1 |
| 80-89 | 4 |
| 70-79 | 2 |

### Summary of Results

| Central Tendency | Variability |
|---|---|
| Mean = 120.43 | Standard deviation = 19.59 |
| Median = 124 | Quartile deviation = 12.83 |
| Mode = 125 | Range = 70 |

Appendix B

Histogram of Total Scores Assigned




Mean = 120.43
Median = 124
Mode = 125

Negatively skewed distribution

Appendix C

## Reliability Data

A. The Instrument

| | |
|---|---|
| Split-half | = .97 |
| Test-retest (repeat) | = .87* |
| Guttman's | = .97 |
| Kuder-Richardson 21 | = .94 |

The standard error of measurement is approximately 4 points

B. Interrater

Median interrater reliability when each standard is analyzed separately = .68

Median interrater reliability when only total scores are considered = .96

*There were only 9 teachers visited by the same teams on two occasions.

Appendix D

## Validities of the Standards

| Standard | Correlation with Total Score |
|---|---|
| 2 (enthusiasm, etc.) | .931 |
| 1 (cognitive clarity) | .919 |
| 8 (diverse learners) | .901 |
| 5 (student interaction) | .900 |
| 7 (objectives/planning) | .888 |
| 6 (positive rapport) | .886 |
| 4 (task oriented) | .874 |
| 9 (professional attitude) | .847 |
| 3 (variety) | .846 |
| 10 (professional growth) | .610 |

In comparing the magnitude of two correlation coefficients, squaring them first and then comparing the resultant coefficients of determination yields a better index of the relative size of the difference.

In other words standard 3 shares over 71 per cent of the variance found in total scores, while standard 10 shares only 37 per cent. Standards 2 and 1 share 87 and 84 per cent, respectively. Thus standard 10 is even lower than it appears.

Appendix E

Team by Team Comparisons

(excluding specialists)

| Team | Mean | Interrater Reliability (all standards) | Interrater Reliability (total scores only) | Average Pearson r |
|---|---|---|---|---|
| alpha* | 104.75 | .82 | .97 | .68 |
| beta | 132.23 | .43 | .75 | .29 |
| gamma | 133 | .58 | .66 | .32 |
| delta | 124.92 | .55 | .96 | .30 |
| omega | 111.42 | .75 | .96 | .58 |

*Alpha differed significantly ($p < .02$) from both beta and gamma. No other means differed significantly from the others.

Note that the two teams with the lowest means had the highest reliabilities.

Appendix F

## Normative Data on Total Scores Assigned

$P_{99}$ = 145.05

$P_{95}$ = 143.88

$P_{90}$ = 140.25

$P_{85}$ = 139.13

$P_{80}$ = 137.17

$P_{75}$ = 136.25

$P_{70}$ = 133.75

$P_{65}$ = 131.13

$P_{60}$ = 130

$P_{55}$ = 128.38

$P_{50}$ = 124

$P_{45}$ = 122.25

$P_{40}$ = 121.5

$P_{35}$ = 116.25

$P_{30}$ = 111.33

$P_{25}$ = 110.58

$P_{20}$ = 106

$P_{15}$ = 94.25

$P_{10}$ = 86

$P_{5}$ = 81.75

$P_{1}$ = 75.22

ANDOVER PUBLIC SCHOOLS
ANDOVER, MASSACHUSETTS

## EFFECTIVE TEACHING PERFORMANCE STANDARDS AND RATING SCALE

### Rationale and Overview

It is generally accepted that effective teaching is the most important element in a good education. The performance standards listed in this booklet suggest those criteria by which effective teaching is appraised in the Andover Public Schools. The criteria are drawn primarily from the research on those factors contributing to teaching effectiveness. (A summary of much of this research is described by Barak Rosenshine and Norma Furst in "The Use of Direct Observation to Study Teaching " in the American Educational Research Association's Second Handbook of Research on Teaching.) These criteria were reviewed in terms of their face validity by teachers and administrators in the Andover Public Schools.

Although each of the standards contributes to effective teaching, a hierarchy is implicit. This hierarchy places the greatest value or emphasis on those standards which relate to the face-to-face interaction of teachers with students. These standards, categorized as instructional implementation include:

1. Exhibits cognitive clarity in presentation.
2. Generates enthusiasm, excitement, and/or interest during presentation.
3. Makes use of variety in presentation.
4. Is task-oriented and exhibits business-like behavior; is able to provide for smooth classroom activity transition and monitor the class doing more than one thing at a time.
5. Provides for student interaction and involvement in the presentation.
6. Creates positive student-teacher rapport.

Effective instructional implementation is dependent on instructional planning. This criterion includes two standards:

7. Selects and generates defensible instructional objectives and plans.
8. Achieves instructional objectives with diverse kinds of learners.

Finally, instructional planning and implementation are affected by a teacher's professional participation. Two standards are included under that criterion:

9. Displays professional attitude.
10. Shows evidence of professional growth.

Following a series of formal and informal observations, the evaluator appraises a teacher's performance based on these standards. Each standard and its related performance activities is reviewed in light of those observations. Overall assessment of performance on a standard is made on a five-point scale ranging from "seldom" to "consistently." A summary evaluation statement follows the rating. This detailed statement describes accomplishments, current inadequacies, efforts made to overcome deficiencies, and suggestions for the future -- all related to a particular standard and its performance activities.

## EFFECTIVE TEACHING PERFORMANCE STANDARDS *

1.0 Exhibits cognitive clarity in presentation.

- 1.1 Displays verbal fluency; able to explain clearly.
- 1.2 Has firm grasp on what is being taught and why.
- 1.3 Is accurate from a factual standpoint.
- 1.4 Displays thorough knowledge of subject matter.
- 1.5 Uses correct language.
- 1.6 Asks a variety of questions which require both "lower cognitive" as well as "higher cognitive response."

2.0 Generates enthusiasm, excitement, and/or interest during presentation.

- 2.1 Exhibits high expectations and assumes personal responsibility for making sure students learn.
- 2.2 Is stimulating and imaginative.
- 2.3 Evidences strong motivation and commitment to teaching.
- 2.4 Exhibits a sense of humor.
- 2.5 Is intellectually stimulating.
- 2.6 Transfers relevant and interesting real life situations into the classroom.
- 2.7 Maintains stimulating and attractive classroom environment.

3.0 Makes use of variety in presentation.

- 3.1 Exhibits sufficient movement, gestures, and variation in voice and eye contact.
- 3.2 Makes effective use of a wide variety of well-selected instructional materials.

*Copyright © 1979 by the Andover (Massachusetts) Public Schools

3.3 Is actively involved in lesson.

3.4 Makes effective use of instructional strategies such as large group, small group, or individualized instruction, independent activities, or contracting.

3.5 Demonstrates appropriate skills.

4.0 Is task-oriented and exhibits business-like behavior; is able to provide for smooth classroom activity transition and monitor the class doing more than one thing at a time.

4.1 Keeps students actively engaged in productive work.

4.2 Displays alertness of what's going on in classroom.

4.3 Has whole group's attention before beginning presentation.

4.4 Uses statements designed to provide an overview and a summary of the lesson.

4.5 Provides opportunity for students to practice the new learning and get feedback.

4.6 Employs a recognized systematic instructional format and pattern.

4.7 Demonstrates purposeful organization of space and materials with regard to the physical appearance as well as high safety and maintenance standards.

5.0 Provides for student interaction and involvement in the presentation.

5.1 Acknowledges, modifies, applies, compares, or summarizes student statements.

5.2 Does not consistently repeat student responses.

5.3 Probes or uses responses which encourage student (or another student) to elaborate upon his or her answers.

5.4 Responds to questions posed by students.

5.5 Responds to substance of student statement rather than form.

5.6 Provides for high frequency of student initiated questions and interaction.

Copyright © 1979 by the Andover (Massachusetts) Public Schools

5.7 Spends time interacting with individual students even during group or whole class lessons.

5.8 Praises or encourages students.

6.0 Creates positive student-teacher rapport.

6.1 Treats students in a manner which reflects an awareness of their developmental and emotional characteristics.

6.2 Shows warmth toward students; has positive attitude toward students.

6.3 Able to handle discipline problems.

6.31 Sets framework for learning/discipline so that students know what to expect; is consistent.

6.32 Is impartial and fair when dealing with students.

6.33 Maintains standards both inside and outside classroom.

6.34 Discusses disciplinary incident with student and follows through on disciplinary action.

6.4 Accepts feelings of students in a non-threatening manner.

6.5 Develops and fosters balance between individual rights and responsibilities.

6.6 Does not appear to get overly excited or scream and holler or use sarcasm at students.

6.7 Interacts positively with students both inside and outside the classroom.

6.8 Students demonstrate a positive attitude toward class and teacher.

6.9 Shows compassion and empathy; sincerely wants students to learn.

Copyright © 1979 by the Andover (Massachusetts) Public Schools

7.0 Selects appropriate and justifiable instructional objectives and generates sound educational plans.

7.1 Exhibits evidence of direct relationship between material covered/used in class and instructional objectives.

7.2 Lessons demonstrate thought and preparation.

7.3 Uses instructional materials in an organized fashion that reflects planning.

7.4 Develops written plans which are based on prescribed curriculum and reflect use of progam manuals, courses of study, guidelines, etc.

7.5 Uses appropriate test data in planning program.

8.0 Achieves instructional objectives with diverse kinds of learners.

8.1 Is able to diagnose individual student's skills, progress and learning style; adapts program accordingly.

8.2 Keeps adequate regularly updated accumulative records; evaluates progress of each student, and adjusts the program accordingly.

8.3 Matches difficulty level of the lesson to ability level of the students through lesson content, questions, and activities.

8.4 Provides material so every student can show achievement within the work period.

8.5 Facilitates the learning for the reluctant learners - brings out students to realize their potential.

8.6 Uses good judgement by being demanding, critical, encouraging, supportive, helpful, or warm, as dictated by the needs of each student in any given situation.

9.0 Displays professional attitude.

9.1 Develops positive relations with colleagues as a team/grade level/department and school staff member.

9.2 Contributes to team/grade level/department and school; willing to share ideas and support them.

9.3 Cooperates with administrative staff.

9.4 Maintains positive working relationship with parents.

Copyright © 1979 by the Andover (Massachusetts) Public Schools

9.5 Participates actively in school activities.

9.6 Provides extra help for students as needed.

9.7 Responds positively to responsible constructive criticism or advice.

10.0 Shows evidence of professional growth.

10.1 Keeps abreast of developments in field through professional reading and active participation in professional associations directly related to field of work.

10.2 Engages in appropriate course work and in-service programs.

10.3 Works with new approaches and materials and is frank in their evaluation.

Copyright © 1979 by the Andover (Massachusetts) Public Schools

## Appendix H

ANDOVER PUBLIC SCHOOLS
ANDOVER, MASSACHUSETTS

EFFECTIVE TEACHING PERFORMANCE RATING SCALE

TEACHER_____ OBSERVER_____ SCHOOL_____ DATE_____

| TO WHAT EXTENT DOES THE TEACHER EXHIBIT THE FOLLOWING STANDARDS: | Seldom | | | | Consistently |
|---|---|---|---|---|---|
| **INSTRUCTIONAL IMPLEMENTATION** | | | | | |
| 1. Exhibits cognitive clarity in presentations: | 1 | 2 | 3 | 4 | 5 |
| Statement:_____ | | | | | |
| 2. Generates enthusiasm, excitement, and/or interest during presentation. | 1 | 2 | 3 | 4 | 5 |
| Statement:_____ | | | | | |

Copyright © 1979 by the Andover (Massachusetts) Public Schools

EFFECTIVE TEACHING PERFORMANCE RATING SCALE (continued)

3. Makes use of variety in presentation. 1 2 3 4 5

Statement: _____

4. Is task-oriented and exhibits business-like behavior; is able to provide for smooth classroom activity transition and monitor the class doing more than one thing at a time. 1 2 3 4 5

Statement: _____

5. Provides for student interaction and involvement in the presentation. 1 2 3 4 5

Statement: _____

Copyright 1979 by the Andover (Massachusetts) Public Schools

EFFECTIVE TEACHING PERFORMANCE RATING SCALE (continued)

6. Creates positive student-teacher rapport. 1 2 3 4 5

Statement: _____

INSTRUCTIONAL PLANNING

7. Selects and generates definable instructional objectives and plans. 1 2 3 4 5

Statement: _____

8. Achieves instructional objectives with diverse kinds of learners. 1 2 3 4 5

Statements: _____

Copyright © 1979 by the Andover (Massachusetts) Public Schools

EFFECTIVE TEACHING PERFORMANCE RATING SCALE (continued)

PROFESSIONAL PARTICIPATION

9. Displays professional attitude. 1 2 3 4 5

Statement: _____

10. Shows evidence of professional growth. 1 2 3 4 5

Statement: _____

I have read this report.

Teacher's Signature _____ Date _____

Teacher's Comments: _____

Copyright © 1979 by the Andover (Massachusetts) Public Schools